High Level Design (HLD)

Credit Card Default Prediction

21/03/2024

Devarshi Choudhury

Contents

Abstract

In an era marked by extraordinary advancements in the financial sector, commercial banks are encountering a formidable challenge: the prediction of credit risk amidst evolving financial landscapes. A pivotal concern among these institutions is the ability to foresee the likelihood of credit default among their clients. This project delves into the realm of predictive analytics, aiming to forecast the probability of credit default based on a comprehensive analysis of credit card owners' characteristics and their payment histories.

Financial threats loom large, painting a narrative of the shifting dynamics within commercial banking. The remarkable progress in the financial industry has brought to light an ever-growing need for accurate risk assessment tools. Among these tools, the ability to anticipate credit defaults stands as a critical frontier for banks aiming to navigate the complexities of modern finance.

Imagine the scenario where a seemingly manageable debt spirals out of control due to unforeseen circumstances such as job loss, medical emergencies, or business downturns. It's a narrative many of us can relate to—the missed credit card payments due to oversight or temporary cash flow challenges. However, what happens when these missed payments persist, stretching into months of financial strain?

This project seeks to address precisely this question by harnessing the power of machine learning algorithms. By analysing a diverse range of data points including demographic information like gender, age, marital status, and behavioural patterns such as previous payments and transaction history, we aim to develop a robust model. This model will not only predict the likelihood of a customer becoming a defaulter but also empower banks with actionable insights to mitigate risks effectively.

In a world where financial stability is paramount, this project endeavours to provide commercial banks with a predictive tool that can potentially redefine their risk management strategies. Through the lens of machine learning, we embark on a journey to enhance the industry's ability to foresee and navigate the intricate landscape of credit default risks.

1. Introduction:

The "Credit Card Default Prediction" project aimed to develop a predictive model to determine the likelihood of a customer defaulting on their credit card payments. The dataset used for this project was the "Default of Credit Card Clients Dataset" obtained from the UCI Machine Learning Repository. This dataset contained information on credit limits, gender, education level, marital status, age, payment history, and bill amounts for 30,000 credit card clients.

The project followed a comprehensive data science pipeline, starting with data exploration and visualization to gain insights into the dataset. Exploratory Data Analysis (EDA) helped in understanding the distributions, correlations, and patterns in the data. Data cleaning and preprocessing steps were performed to handle missing values, correct data types, and engineer new features like the Credit Utilization Ratio.

Several machine learning models were trained and evaluated, including Random Forest, LightGBM, and XGBoost, to predict credit card defaults. The models were assessed based on their performance metrics such as ROC-AUC score, accuracy, precision, recall, and F1-score. The best-performing model was selected for deployment.

The web application was built using Streamlit, allowing users to input their demographic and payment information. Upon submission, the model predicted whether the customer was likely to default on their credit card payments. The application displayed the prediction results, including the Credit Utilization Ratio and the Probability of Default.

Overall, this project demonstrated the process of developing a credit card default prediction model, from data exploration and preprocessing to model training and deployment. The web application provides a user-friendly interface for users to assess their credit risk and make informed financial decisions.

1.1 Why this High-Level Document?

The High-Level Design (HLD) document serves the purpose of elaborating on the project's structure, employing non-technical to mildly technical terms that are understandable to system administrators. Its objectives are to enhance the current project description with the necessary details to serve as a suitable model for coding. Additionally, this document aims to identify any potential contradictions before the coding phase and can act as a reference manual illustrating how modules interact at a higher level.

The HLD will:

1. Present all of design aspects and define them in detail.

2. Describe the user interface being implemented.

3. Describe software interfaces.

4. Include Design features and architecture of the project.

1.2 Scope

The High-Level Design (HLD) document outlines the complete project structure in various sections, including data ingestion, data pre-processing, solution development, and deployment. Each section is described using non-technical to mildly technical terms, ensuring clarity for system administrators.

1.3 Definitions

Term	Description
IDE	Integrated Development Environment
EDA	Exploratory Data Analysis
VS Code	Visual Studio Code
LightGBM	Light Gradient Boosting Machine

2. General Description:

2.1 Product Perspective

The Credit Card Default Prediction is a Machine Learning model based on the XGBoost algorithm, designed to predict whether a customer will default on their next month's payment. Additionally, the app provides a probability estimation of default.

2.2 Problem Statement:

The financial landscape is witnessing a noticeable trend in the credit risk faced by commercial banks due to remarkable advancements in the financial industry. Among the significant challenges encountered by commercial banks is the accurate prediction of credit clients' risk. The objective is to forecast the likelihood of credit default by leveraging the characteristics and payment history of credit card owners.

2.3 Proposed Solution:

The proposed solution is a web application that serves as an intuitive interface for users to input customer details. These details are then processed by a trained machine learning model running in the backend. The model predicts the likelihood of a customer defaulting on their credit card payments. The web application displays both the prediction outcome and the corresponding probability of default on the user-facing frontend page, providing users with valuable insights for decision-making.

2.4 Technical Requirements:

The machine learning model was developed using Python 3.10.12 and essential libraries such as Scikit-learn, Pandas, and XGBoost. These libraries were crucial for accurate credit card default prediction, feature engineering, and model evaluation.

The web application's frontend was created using Streamlit, a user-friendly library for building interactive web apps with Python. Streamlit allowed for the seamless integration of user input forms and the display of prediction results in real-time.

The project utilized Google Colab for model training and development, ensuring efficient collaboration and access to GPU resources for faster computations.

Deployment was achieved using platforms like Streamlit Sharing, allowing the web application to be accessible via a URL link for end-users.

3. Dataset Information:

Data Source: Default of Credit Card Clients Dataset

A total of 25 variables have been identified:

ID: ID of each client

LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary = credit)

SEX: Gender (1=male, 2=female)

EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)

MARRIAGE: Marital status (1=married, 2=single, 3=others)

AGE: Age in years

PAY_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for

one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)

PAY_2: Repayment status in August, 2005 (scale same as above)

PAY_3: Repayment status in July, 2005 (scale same as above)

PAY_4: Repayment status in June, 2005 (scale same as above)

PAY_5: Repayment status in May, 2005 (scale same as above)

PAY_6: Repayment status in April, 2005 (scale same as above)

BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)

BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)

BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)

BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)

BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)

BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)

PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)

PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)

PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)

PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)

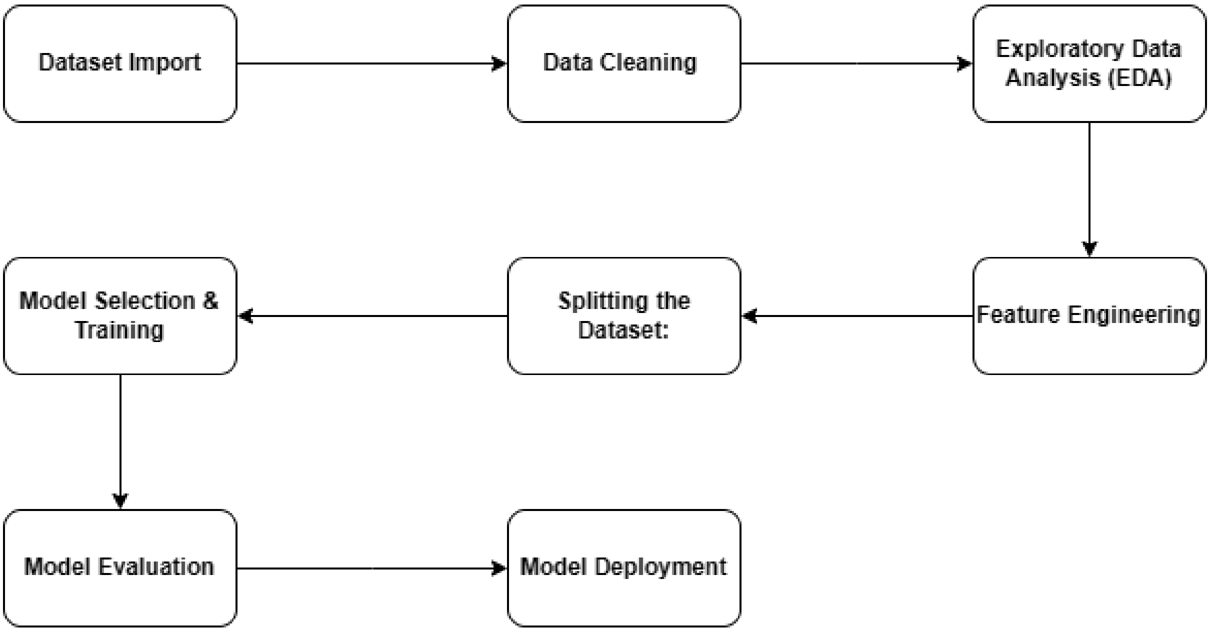
PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)

PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)

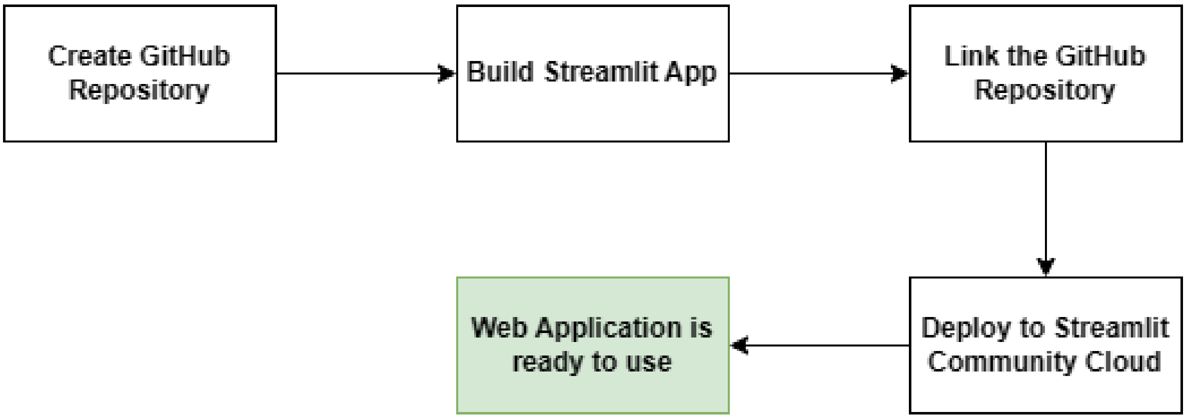
default.payment.next.month: Default payment (1=yes, 0=no)

4. Design Details

4.1 Process flow:



4.2 Deployment Process:


Create GitHub Repository
Build Streamlit App
Link the GitHub Repository
Deploy to Streamlit Community Cloud
Web Application is ready to use

5. Tools and Technologies used

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn are used to build such a model

- Google Colab was used as the work environment
- For Visualization of the plots, Matplotlib and Seaborn are used.
- VS Code was used to build app.py file for web application.
- GitHub is used as Version Management System.
- Streamlit for cloud deployment.
- Statsmodels library is used for checking relationship with target variable.

5.1 Reusability

The written code and utilized components are designed for easy reuse. Additionally, Model Pickle files are generated for each tested model, facilitating their reuse as well.

5.2 Application Compatibility

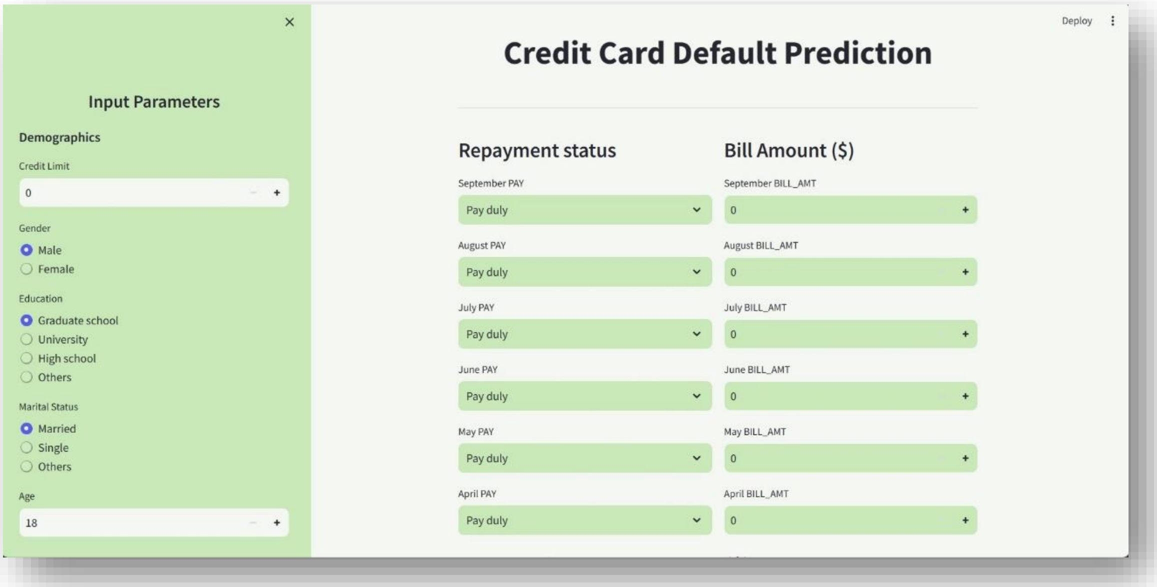
The various components or modules of this project interface with each other using Python version 3.10.12.

6. Deployment:

The application has been deployed on the web using the Streamlit Community Cloud platform.

Link: https://credit-card-default-prediction-lh653hvkh6akelqbfzzf5g.streamlit.app/

6.1 User Interface:



Input Parameters

Demographics

Credit Limit: 0

Gender
● Male
○ Female

Education
● Graduate school
○ University
○ High school
○ Others

Marital Status
● Married
○ Single
○ Others

Age: 18

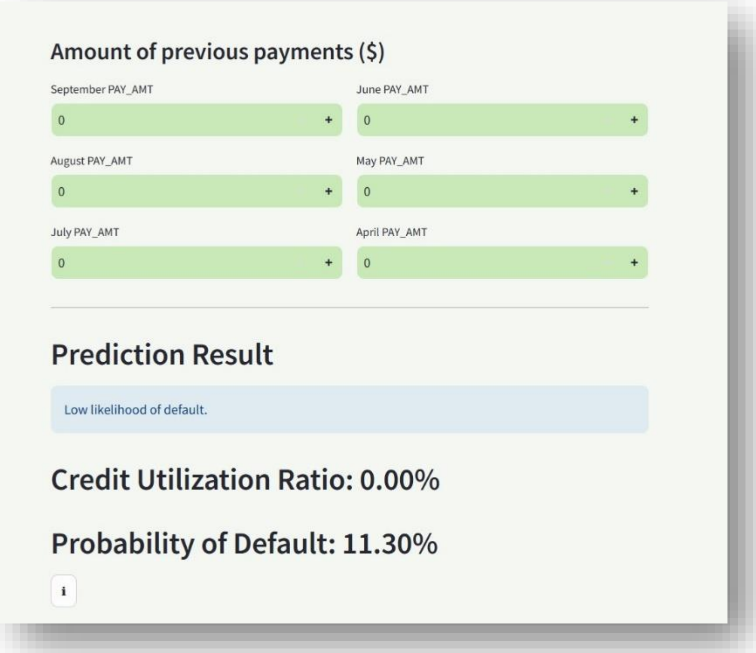
Credit Card Default Prediction

Repayment status

September PAY: Pay duly
August PAY: Pay duly
July PAY: Pay duly
June PAY: Pay duly
May PAY: Pay duly
April PAY: Pay duly

Bill Amount ($)

September BILL_AMT: 0
August BILL_AMT: 0
July BILL_AMT: 0
June BILL_AMT: 0
May BILL_AMT: 0
April BILL_AMT: 0



Amount of previous payments ($)

September PAY_AMT: 0
August PAY_AMT: 0
July PAY_AMT: 0
June PAY_AMT: 0
May PAY_AMT: 0
April PAY_AMT: 0

Prediction Result

Low likelihood of default.

Credit Utilization Ratio: 0.00%

Probability of Default: 11.30%

7. Conclusion:

The Credit Card Default Prediction model developed in this project offers a proactive approach to managing credit risks for financial institutions. Leveraging machine learning algorithms like Random Forest, LightGBM, and XGBoost, the model forecasts the likelihood of a customer defaulting on credit card payments. This predictive capability provides valuable insights, enabling institutions to take timely actions such as offering tailored payment plans or limiting credit exposure to prevent defaults and minimize financial losses.

By integrating various customer credit card usage patterns and payment history, the credit card default prediction model enhances the credit assessment process, offering a more comprehensive understanding of individual financial behaviours. This model not only empowers institutions with foresight but also helps strengthen customer relationships and optimize credit management strategies. Ultimately, the credit card default prediction model facilitates smarter decision-making, improving business profitability and resilience in the ever-evolving landscape of the financial industry.